GroupGazer: A Tool to Compute the Gaze per Participant in Groups with Integrated Calibration to Map the Gaze Online to a Screen or Beamer Projection

Wolfgang Fuhl[1,*], Daniel Weber[1] and Shahram Eivazi[1,2]
[1]*University Tübingen, Sand 14, Tübingen, Germany*
[2]*FESTO, Ruiter Str. 82, Esslingen am Neckar, Germany*

Keywords: Eye Tracking, Gaze, Gaze Group, Calibration, Group Calibration, Gaze Mapping.

Abstract: In this paper we present GroupGaze. It is a tool that can be used to calculate the gaze direction and the gaze position of whole groups. GroupGazer calculates the gaze direction of every single person in the image and allows to map these gaze vectors to a projection like a projector. In addition to the person-specific gaze direction, the person affiliation of each gaze vector is stored based on the position in the image. Also, it is possible to save the group attention after a calibration. The software is free to use and requires a simple webcam as well as an NVIDIA GPU.

1 INTRODUCTION

Eye tracking is an important input modality and information source in the modern world (Cognolato et al., 2018). In the field of human-machine interaction, the gaze signal is used and further researched for interaction with robots (Willemse and Wykowska, 2019) but also other technical devices (Wanluk et al., 2016). This involves not only simple control but also collaboration in which a human communicates complex behavior to a robot or system (Palinko et al., 2016). Interaction with the eyes is also an interesting source of information in the field of computer games (Alkan and Cagiltay, 2007). Eye interaction is many times faster than mouse interaction, which could revolutionize the professional computer gaming field (Jönsson, 2005). In the field of virtual reality, gaze information can be used to render only small areas of the scene in high resolution, leading to a significant reduction in the resources consumption of the devices (Meng et al., 2018). Another important area in which the gaze signal plays an important role is driver observation. Here it is necessary to assess whether the driver is able to control the vehicle or is too tired in the case of autonomous driving to take over the vehicle (Zandi et al., 2019). Of course, this also applies to car rental companies, for which it is important to know whether the driver is, for example, intoxicated or an unsafe driver (Maurage et al., 2020). In the field of medicine, research is also being conducted into methods of self-diagnosis (Clark et al., 2019). This involves, for example, the early detection of Alzheimer's disease (Crawford, 2015), strokes (Matsumoto et al., 2011), as well as eye defects (Eide et al., 2019) or autism (Boraston and Blakemore, 2007). In the field of safety, the eye signal also gains increasingly more interest (Katsini et al., 2020; Fuhl et al., 2021a). This is due to the fact that personal behavior is reflected in the gaze signal, which can be used to identify the person (Fuhl et al., 2021b). Other information contained in the eye is the cognitive load based on pupil dilation (Chauliac et al., 2020), attention (Chita-Tegmark, 2016), procedural strategies (Jenke et al., 2021) and many others. A relatively new area in which the eye tracking signal is used is behavioral research (Yang and Krajbich, 2021; Das et al., 2018). Here it is on the one hand about extracting expert knowledge from the eye signal and passing this knowledge to trainees (Manning et al., 2003; Hoghooghi et al., 2020). This concerns, all areas in which the training is only possible with expensive tools and training devices (Vijayan et al., 2018). In the area of medicine the main interest is to distill the expert knowledge better (Quen et al., 2021; Manning et al., 2003). Another area of behavioral research which is also the subject of this thesis is group behavior (Hwang and Lee, 2020; Reichenberger et al.,

*Corresponding author

Fuhl, W., Weber, D. and Eivazi, S.
GroupGazer: A Tool to Compute the Gaze per Participant in Groups with Integrated Calibration to Map the Gaze Online to a Screen or Beamer Projection.
DOI: 10.5220/0011607000003417
In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023) - Volume 2: HUCAPP*, pages 109-116
ISBN: 978-989-758-634-7; ISSN: 2184-4321

Figure 1: Detection results of the proposed tool GroupGazer on a group photo. **The image has a high resolution in the pdf so you can zoom in to see everything.** A larger version of the image with more people is in the supplementary material. The image is taken from www.pexels.com.

2020; Kredel et al., 2017). Here there is research in the area of teaching (Korbach et al., 2020; Schneider et al., 2008; Jarodzka et al., 2020) but also in dynamic environments like sports (Oldham et al., 2021; Du Toit et al., 2009; Reneker et al., 2020).

The current problems in the field of behavioral research for groups, is that there is no freely available software for this. Therefore, research groups have to resort to expensive solutions such as multiple worn eye trackers. This creates further issues like the assignment of the important areas between the different scene cameras. One way around this is to use virtual reality together with eye tracking. However, this also changes the behavior of the test persons and cannot be carried out over longer periods of time with regard to motion sickness. Alternatively to worn eye trackers, there is also the possibility to use external cameras. In this case, the researchers have to implement their technical solutions independently, which often leads to dependencies on other working groups and is also an expensive undertaking due to the image processing cameras which are usually used.

In this paper, we present software that allows anyone to use a simple webcam for gaze estimation of groups and calibration each subject in parallel. By doing so, we hope to enable anyone to conduct behavioral group research. Our contributions to the state of the art are:

1. A tool to record the gaze of groups and calibrate each individual in parallel.
2. The tool has no specialized hardware requirements and only needs an NVIDIA GPU with at least 4 GB memory (We used a 1050 ti with 4 GB).
3. Stores the gaze per person as well as the average gaze location of the group.

2 RELATED WORK

Since our software is the combination of several research fields, we have divided the related work into three categories. The first category is face recognition, the second category is appearance based gaze estimation, and the third category is gaze based group behavior research.

2.1 Face Detection

Face recognition in arbitrary environments is still a very challenging field of research. Here, an arbitrarily large image is given, and all faces must be detected. This often involves occlusions, different head positions, changes in lighting conditions, and of course the faces in the image have different resolutions. The first very successful approach was presented by Viola and Jones (Viola and Jones, 2004). This is based on hair features and trained using AdaBoost. The next major step was achieved with deformable part models (DPM) (Yan et al., 2014). Compared to feature-based approaches, DPM is much more robust but requires significantly more computational effort. With the advent of deep neural networks, however, the state of the art was again significantly improved (Bai et al., 2018; Jiang and Learned-Miller, 2017; Li et al., 2019; Zhang et al., 2020). The first extension of neural networks was the combination of face detection with face matching (Zhang and Zhang, 2014; Zhang et al., 2016). Current methods for face detection follow two directions. The first direction is the multistage approach, which is based on a region proposal neural network followed by validation of the proposed faces. The most notable representatives of this direction of development are RCNN (Girshick et al., 2014), almost RCNN (Girshick, 2015), and faster RCNN (Ren et al., 2015). The second direction of development is direct methods such as single shot multibox detector (SSD) (Liu et al., 2016) or you only look once (YOLO) (Redmon et al., 2016). For YOLO, there are

already multiple versions, which consume even less resources at approximately the same detection rate. The advantage of the direct approaches, is the faster execution and the smaller resource consumption. The multilayer methods, on the other hand, provide a better detection rate and fewer misclassifications.

2.2 Appearance Based Gaze Estimation

Here, the entire facial image or eye area of a person is used to directly determine the gaze vector via a neural network. The first work in this area is from 1994 (Baluja and Pomerleau, 1994) and was extended in (Tan et al., 2002) by linear projection functions. These methods require very expensive calibration, since the neural network was trained for each person individually with many training examples. The first extensions to reduce the effort in calibration were a Gaussian process regression (Williams et al., 2006), saliency maps (Sugano et al., 2012), and optimal selection using a linear regression (Lu et al., 2014). While all of these methods advanced the state of the art, the appearance based approach still had many limitations, such as a fixed head position and per-person calibration. With deep neural networks and the advent of big data, this has changed significantly. In (Zhang et al., 2015) the first successful approach was presented, which realized appearance based gaze estimation with deep neural networks. The first extensions used, in addition to eye images, the subjects' faces, which resulted in a significant improvement (Krafka et al., 2016; Kellnhofer et al., 2019). For extreme head positions and strongly deviating gaze angles to the head orientation, an asymmetric regression was presented (Cheng et al., 2020).

2.3 Gaze Based Group Behavior Research

In this section, we would like to mention and briefly explain only some works from this area, since our software is made for this purpose but does not perform a behavior research study.

The first area in behavioral research which can also be applied to groups is mind wandering (Hutt et al., 2017; Hutt et al., 2019). Mind wandering is a shift in attention to task-unrelated thoughts. This is an interesting effect for teaching since it negatively influences the learning performance of students (Robertson et al., 1997; Smallwood et al., 2008; Hutt et al., 2019). Mind wandering itself is a special form of disengagement and has to be separated from boredom or off-task behaviors (Cocea and Weibelzahl, 2010; Mills et al., 2014; Hutt et al., 2019). Another interesting social behavior is gaze following (Aung et al., 2018). This gaze following is a form of communication and socializing. In some scenarios it has to be done only for a single person (Judd et al., 2009; Fathi et al., 2012) but in modern research entire scenes with multiple persons are evaluated and analyzed (Aung et al., 2018; Mukherjee and Robertson, 2015; Marin-Jimenez et al., 2014; Recasens et al., 2017; Recasens, 2016). Nowadays, psychologists use behavior observation methods in classrooms as well as direct behavior ratings (Woolverton and Pollastri, 2021). While both methods are valid and also used by teachers themselves, they are limited in effectiveness due to the attentional limits of the human observers as well as their induced biases (McIntyre and Foulsham, 2018). Modern research focuses on establishing intelligent classroom technologies with eye tracking and voice recording (Woolverton and Pollastri, 2021; McIntyre and Foulsham, 2018). Those methods have their limitations due to the data security but deliver more insights and allow reducing the induced bias by humans (Woolverton and Pollastri, 2021; McIntyre and Foulsham, 2018; McParland et al., 2021).

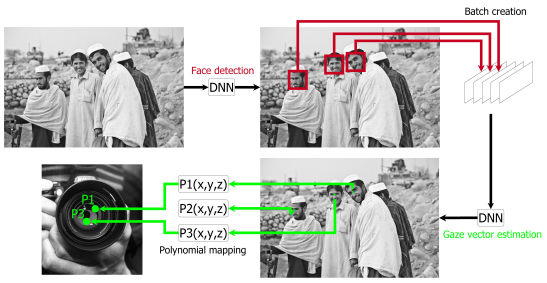

Figure 2: The workflow of our approach. We first detect the faces and compute the gaze vector using an appearance based approach. Each person is calibrated using mouse clicks on a projection in parallel. The fitted polynomials are afterwards used to map the gaze to the projection.

3 METHOD

Figure 2 shows the workflow of our approach. GroupGazer first opens a video stream on an available camera. Afterwards, all faces in the image are detected. If not all desired faces are detected, GroupGazer offers an upscaling factor, which can be set by the user. This upscaling factor resizes the input image to allow the face detection to detect even very small faces in the image. After the face detection, all detected faces are extracted from the image and resized to 100×100 pixels in a gray scale image. These images are grouped together to form a batch which is given to the gaze vector estimation DNN.

Table 1: Shows the architecture of our face detection deep neural network. The architecture is copied from dlib (King, 2009) and uses the max margin (King, 2015) training procedure. We modified the model in terms of tensor normalization (Fuhl, 2021b) and gradient centralization (Fuhl and Kasneci, 2021) as well as convolution size and depth.

Level	Gaze estimator
Input	RGB image any resolution
1	Pyramid layer with six stages
2	5×5 Conv, dep 8, 2×2 down, BN, ReLu, TN
3	3×3 Conv, dep 8, 2×2 down, BN, ReLu, TN
4	3×3 Conv, dep 8, 2×2 down, BN, ReLu, TN
5	5×5 Conv, dep 16, BN, ReLu, TN
6	3×3 Conv, dep 16, BN, ReLu, TN
7	3×3 Conv, dep 16, BN, ReLu, TN
8	7×7 Conv, dep 1

Table 2: Shows the architecture of our gaze estimation deep neural network. It has the structure of a ResNet-34 (He et al., 2016) and uses the leaky maximum propagation blocks (Fuhl, 2021a), tensor normalization (Fuhl, 2021b), as well as the weight and gradient centralization (Fuhl and Kasneci, 2021).

Level	Gaze estimator
Input	Gray scale image 100×100
1	5×5 Conv, dep 32
2	ReLu with tensor normalization
3	2×2 Max pooling
4	3 Max blocks, 2×2 d, 3×3 C, dep 64, BN
5	ReLu with tensor norm
6	3 Max blocks, 2×2 d, 3×3 C, dep 128, BN
7	ReLu with tensor norm
8	3 Max blocks, 2×2 d, 3×3 C, dep 256, BN
9	ReLu with tensor norm
10	Fully connected, 512 outputs
11	ReLu
12	Fully, 7 (3,7 for validation))

The batch size can also be set by the user. This fixed batch size allows GroupGazer to have a static runtime and if there are fewer faces in the image, the rest of the batch is filled with black images. GroupGazer can be used with a 1050 ti graphics card for up to 40 faces in real time, which is also dependent on the input resolution to the face detection DNN. For newer GPUs more faces can be set by the user as well as larger input image resolutions for the face detection. The gaze estimation DNN processes the entire batch and computes a starting position (First two values), an accuracy of the starting position (Third value), the gaze vector (Forth to sixth value), as well as an accuracy of the gaze vector (Seventh value). With this information, each face has a gaze vector and an estimated accuracy. With the gaze vector and the starting position, a polynomial is used to map the gaze vector to a projection or monitor. The degree of the polynomial can be specified by the user, and the calibration procedure works as follows. The teacher or adviser tells the students to look at his mouse cursor position. On a left mouse click, all gaze vectors which are seen as valid and accurate are stored together with the click location. This is repeated multiple times. Afterwards, for each user, the polynomial is fitted in the least squares sense. With those polynomials, the mapping and therefore the gaze location is computed for each user. The reidentification of users is done by the smallest euclidean distance to the last detections, and the new position is not allowed to leaf the last face detection bounding box. This is a simple procedure but saves a lot of computational resources since no additional network has to be used. In addition, it is much more robust since fine-tuning a Network online usually needs multiple examples to deliver reliable results, even if we use the hypersphere approach (Xie et al., 2019) or siam networks (Abdelpakey and Shehata, 2019).

The used model architectures can be seen in Table 1 and 2. Our face detection model is similar to the model from dlib (King, 2009) we only made some slight changes which improve the accuracy of the model and only impact the runtime slightly. For gaze estimation we used the architecture of a ResNet-34 (He et al., 2016) since it has a good accuracy and is resource saving in contrast to the other networks. We modified the ResNet-34 architecture only by adding some novel normalization (Fuhl, 2021b; Fuhl and Kasneci, 2021), the landmark validation loss (Fuhl and Kasneci, 2019), as well as leaky maximum propagations instead of the residual connections (Fuhl, 2021a).

4 EVALUATION

Gaze360 (Kellnhofer et al., 2019) is a huge data set with 3D gaze annotations recorded using multiple cameras covering 360 degree. The recordings were conducted indoor and outdoor with 238 subjects. The dataset contains large head variations as well as distances of the subjects to the camera. We only used approximately 80,000 images of this data set since the data set contains also human heads from behind as well as some partially covered heads which we removed from our data for training and evaluation. The train and test split was done by randomly selecting 20% for testing and 80% for training.

DLIB (King, 2009) data set contains images of various resolutions. Each image can have multiple faces which are annotated with bounding boxes. In

Table 3: Face detection results on DLIB data set (King, 2009) with percition and recall. We compare our model to other approaches in therms of detection percentage as well as runtime in milliseconds (ms) for one hundred images in average. OoM means out of memory exception.

Method	Percision	Recall	Runtime GPU (ms)	
			1920×1280	300×300
Proposed	0,99	0,89	67	3
dlib (King, 2009)	0,99	0,88	175	8
Res-34 & Faster-RCNN (Ren et al., 2015)	0,99	0,91	OoM	22 & 1
Yolov5s (Redmon et al., 2016)	0,99	0,89	OoM	10

Table 4: Appearance based gaze estimation results on the Gaze360 (Kellnhofer et al., 2019) dataset. We compared our model to other approaches and evaluated the gaze start estimation in average euclidean distance in pixel as well as the gaze vector estimation in degree. Time is measured for one face image as average over one thousand.

Method	Gaze start	Gaze vector	Runtime GPU (ms)
Proposed	0,6	0,2	3
ResNet-34 (He et al., 2016)	0,9	0,5	8
ResNet-50 (He et al., 2016)	0,5	0,2	12
MobileNet (Howard et al., 2017)	1,8	1,6	7
MobileNetv2 (Sandler et al., 2018)	1,7	1,6	7

total the data set contains 7213 images and 11480 annotated faces. We made a random 50% to 50% split and used the first half for training and the second half for evaluation (One image more for training due to the uneven number). The images in the dataset are taken from other public data sets and annotated by the authors of (King, 2009).

Table 5: Accuracy of the proposed tool for different distances to the camera. The results are the average accuracy over three subjects on a TV screen with a diagonal of 108cm.

Distance subjects	1m	2m	3m	4m	5m	6m
Average Error	4cm	5cm	8cm	12cm	15cm	19cm

In Table 3 and 4 our models are compared with other approaches. For face detection (Table 3), it can be seen that we have chosen a tradeoff between detection rate and runtime. The recognition rate of our approach can be further increased via the upscaling factor. However, this also increases the computation time, which also increases the runtime per image. For Yolo this is not possible, because the memory usage for images larger than 300 becomes too large. For the backbone of the faster-RCNN, the memory consumption is also too high for a large resolution. Which is also the main reason why we decided against YOLO and the faster-RCNN. In addition, both the faster-RCNN with backbone and the YOLO need a fixed input resolution with which they have to be trained. For our fully convolutional approach inspired by the dlib architecture, this is not necessary.

For the gaze direction determination, you can clearly see that our net runs significantly faster than the other nets. This is due to the fact that our layers use less depth than, for example, ResNet-34. The MobileNets cannot show their advantage on the GPU, since they cause cache conflicts here, whereby parts of the code are executed serialized. On a CPU, MobileNet would be significantly faster than our net, but with about 160 ms per face too slow for a real-time evaluation. In terms of results, ResNet-50 is the most accurate, closely followed by our network. In addition to accuracy, if we consider runtime on a GPU, our network is clearly ahead, which is why we chose our architecture.

5 CONCLUSION

In this paper, we have presented GroupGazer. This is a software that allows to determine the gaze direction of groups per person. This gaze determination is done online on a conventional computer with an NVIDIA GPU. GroupGazer allows each person in the group to be calibrated in parallel so that the individual gaze vectors can be mapped to a projection, such as that of a projector or large monitor. The software is intended to support behavioral research and thus make it possible to easily record the gaze positions of groups.

ACKNOWLEDGEMENTS

Daniel Weber is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645

REFERENCES

Abdelpakey, M. H. and Shehata, M. S. (2019). Dp-siam: Dynamic policy siamese network for robust object tracking. *IEEE Transactions on Image Processing*, 29:1479–1492.

Alkan, S. and Cagiltay, K. (2007). Studying computer game learning experience through eye tracking. *British Journal of Educational Technology*, 38(3):538–542.

Aung, A. M., Ramakrishnan, A., and Whitehill, J. R. (2018). Who are they looking at? automatic eye gaze following for classroom observation video analysis. *International Educational Data Mining Society*.

Bai, Y., Zhang, Y., Ding, M., and Ghanem, B. (2018). Finding tiny faces in the wild with generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–30.

Baluja, S. and Pomerleau, D. (1994). Non-intrusive gaze tracking using artificial neural networks. Technical report, CARNEGIE-MELLON UNIV PITTSBURGH PA DEPT OF COMPUTER SCIENCE.

Boraston, Z. and Blakemore, S.-J. (2007). The application of eye-tracking technology in the study of autism. *The Journal of physiology*, 581(3):893–898.

Chauliac, M., Catrysse, L., Gijbels, D., and Donche, V. (2020). It is all in the" surv-eye": Can eye tracking data shed light on the internal consistency in self-report questionnaires on cognitive processing strategies?. *Frontline Learning Research*, 8(3):26–39.

Cheng, Y., Zhang, X., Lu, F., and Sato, Y. (2020). Gaze estimation by exploring two-eye asymmetry. *IEEE Transactions on Image Processing*, 29:5259–5272.

Chita-Tegmark, M. (2016). Social attention in asd: A review and meta-analysis of eye-tracking studies. *Research in developmental disabilities*, 48:79–93.

Clark, R., Blundell, J., Dunn, M. J., Erichsen, J. T., Giardini, M. E., Gottlob, I., Harris, C., Lee, H., Mcilreavy, L., Olson, A., et al. (2019). The potential and value of objective eye tracking in the ophthalmology clinic. *Eye*, 33(8):1200–1202.

Cocea, M. and Weibelzahl, S. (2010). Disengagement detection in online learning: Validation studies and perspectives. *IEEE transactions on learning technologies*, 4(2):114–124.

Cognolato, M., Atzori, M., and Müller, H. (2018). Head-mounted eye gaze tracking devices: An overview of modern devices and recent advances. *Journal of rehabilitation and assistive technologies engineering*, 5:2055668318773991.

Crawford, T. J. (2015). The disengagement of visual attention in alzheimer's disease: a longitudinal eye-tracking study. *Frontiers in aging neuroscience*, 7:118.

Das, M., Ester, P., and Kaczmirek, L. (2018). *Social and behavioral research and the internet: Advances in applied methods and research strategies*. Routledge.

Du Toit, P. J., Kruger, P. E., Chamane, N., Campher, J., and Crafford, D. (2009). Sport vision assessment in soccer players and sport science. *African Journal for Physical Health Education, Recreation and Dance*, 15(4):594–604.

Eide, M. G., Watanabe, R., Heldal, I., Helgesen, C., Geitung, A., and Soleim, H. (2019). Detecting oculomotor problems using eye tracking: Comparing eyex and tx300. In *2019 10th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, pages 381–388. IEEE.

Fathi, A., Hodgins, J. K., and Rehg, J. M. (2012). Social interactions: A first-person perspective. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1226–1233. IEEE.

Fuhl, W. (2021a). Maximum and leaky maximum propagation. *arXiv preprint arXiv:2105.10277*.

Fuhl, W. (2021b). Tensor normalization and full distribution training. *arXiv preprint arXiv:2109.02345*.

Fuhl, W., Bozkir, E., and Kasneci, E. (2021a). Reinforcement learning for the privacy preservation and manipulation of eye tracking data. In *Proceedings of IEEE International Joint Conference on Neural Networks*.

Fuhl, W. and Kasneci, E. (2019). Learning to validate the quality of detected landmarks. In *International Conference on Machine Vision, ICMV*.

Fuhl, W. and Kasneci, E. (2021). Weight and gradient centralization in deep neural networks. In *Proceedings of IEEE International Joint Conference on Neural Networks*.

Fuhl, W., Sanamrad, N., and Kasneci, E. (2021b). The gaze and mouse signal as additional source for user fingerprints in browser applications. *arXiv preprint arXiv:2101.03793*.

Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Hoghooghi, S., Popovic, V., and Swann, L. (2020). Novice to expert real-time knowledge transition in the context of x-ray airport security. In *Proceedings of DRS 2020 International Conference: Synergy. Vol. 4*. Design Research Society (UK).

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Hutt, S., Krasich, K., Mills, C., Bosch, N., White, S., Brockmole, J. R., and D'Mello, S. K. (2019). Automated gaze-based mind wandering detection during computerized learning in classrooms. *User Modeling and User-Adapted Interaction*, 29(4):821–867.

Hutt, S., Mills, C., Bosch, N., Krasich, K., Brockmole, J., and D'mello, S. (2017). " out of the fr-eye-ing pan"

towards gaze-based models of attention during learning with technology in the classroom. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, pages 94–103.

Hwang, Y. M. and Lee, K. C. (2020). An eye-tracking paradigm to explore the effect of online consumers' emotion on their visual behaviour between desktop screen and mobile screen. *Behaviour & Information Technology*, pages 1–12.

Jarodzka, H., Skuballa, I., and Gruber, H. (2020). Eye-tracking in educational practice: Investigating visual perception underlying teaching and learning in the classroom. *Educational Psychology Review*, pages 1–10.

Jenke, L., Bansak, K., Hainmueller, J., and Hangartner, D. (2021). Using eye-tracking to understand decision-making in conjoint experiments. *Political Analysis*, 29(1):75–101.

Jiang, H. and Learned-Miller, E. (2017). Face detection with the faster r-cnn. In *2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017)*, pages 650–657. IEEE.

Jönsson, E. (2005). If looks could kill–an evaluation of eye tracking in computer games. *Unpublished Master's Thesis, Royal Institute of Technology (KTH), Stockholm, Sweden*.

Judd, T., Ehinger, K., Durand, F., and Torralba, A. (2009). Learning to predict where humans look. In *2009 IEEE 12th international conference on computer vision*, pages 2106–2113. IEEE.

Katsini, C., Abdrabou, Y., Raptis, G. E., Khamis, M., and Alt, F. (2020). The role of eye gaze in security and privacy applications: Survey and future hci research directions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–21.

Kellnhofer, P., Recasens, A., Stent, S., Matusik, W., and Torralba, A. (2019). Gaze360: Physically unconstrained gaze estimation in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6912–6921.

King, D. E. (2009). Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758.

King, D. E. (2015). Max-margin object detection. *arXiv preprint arXiv:1502.00046*.

Korbach, A., Ginns, P., Brünken, R., and Park, B. (2020). Should learners use their hands for learning? results from an eye-tracking study. *Journal of Computer Assisted Learning*, 36(1):102–113.

Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., and Torralba, A. (2016). Eye tracking for everyone. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2176–2184.

Kredel, R., Vater, C., Klostermann, A., and Hossner, E.-J. (2017). Eye-tracking technology and the dynamics of natural gaze behavior in sports: A systematic review of 40 years of research. *Frontiers in psychology*, 8:1845.

Li, Z., Tang, X., Han, J., Liu, J., and He, R. (2019). Pyramidbox++: high performance detector for finding tiny face. *arXiv preprint arXiv:1904.00386*.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. (2016). Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer.

Lu, F., Sugano, Y., Okabe, T., and Sato, Y. (2014). Adaptive linear regression for appearance-based gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 36(10):2033–2046.

Manning, D., Ethell, S. C., and Crawford, T. (2003). Eye-tracking afroc study of the influence of experience and training on chest x-ray interpretation. In *Medical Imaging 2003: Image Perception, Observer Performance, and Technology Assessment*, volume 5034, pages 257–266. International Society for Optics and Photonics.

Marin-Jimenez, M. J., Zisserman, A., Eichner, M., and Ferrari, V. (2014). Detecting people looking at each other in videos. *International Journal of Computer Vision*, 106(3):282–296.

Matsumoto, H., Terao, Y., Yugeta, A., Fukuda, H., Emoto, M., Furubayashi, T., Okano, T., Hanajima, R., and Ugawa, Y. (2011). Where do neurologists look when viewing brain ct images? an eye-tracking study involving stroke cases. *PloS one*, 6(12):e28928.

Maurage, P., Masson, N., Bollen, Z., and D'Hondt, F. (2020). Eye tracking correlates of acute alcohol consumption: A systematic and critical review. *Neuroscience & Biobehavioral Reviews*, 108:400–422.

McIntyre, N. A. and Foulsham, T. (2018). Scanpath analysis of expertise and culture in teacher gaze in real-world classrooms. *Instructional Science*, 46(3):435–455.

McParland, A., Gallagher, S., and Keenan, M. (2021). Investigating gaze behaviour of children diagnosed with autism spectrum disorders in a classroom setting. *Journal of Autism and Developmental Disorders*, pages 1–16.

Meng, X., Du, R., Zwicker, M., and Varshney, A. (2018). Kernel foveated rendering. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 1(1):1–20.

Mills, C., Bosch, N., Graesser, A., and D'Mello, S. (2014). To quit or not to quit: predicting future behavioral disengagement from reading patterns. In *International Conference on Intelligent Tutoring Systems*, pages 19–28. Springer.

Mukherjee, S. S. and Robertson, N. M. (2015). Deep head pose: Gaze-direction estimation in multimodal video. *IEEE Transactions on Multimedia*, 17(11):2094–2107.

Oldham, J. R., Master, C. L., Walker, G. A., Meehan III, W. P., and Howell, D. R. (2021). The association between baseline eye tracking performance and concussion assessments in high school football players. *Optometry and Vision Science*, 98(7):826–832.

Palinko, O., Rea, F., Sandini, G., and Sciutti, A. (2016). Robot reading human gaze: Why eye tracking is bet-

ter than head tracking for human-robot collaboration. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5048–5054. IEEE.

Quen, M. T. Z., Mountstephens, J., Teh, Y. G., and Teo, J. (2021). Medical image interpretation training with a low-cost eye tracking and feedback system: A preliminary study. *Healthcare Technology Letters*.

Recasens, A., Vondrick, C., Khosla, A., and Torralba, A. (2017). Following gaze in video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1435–1443.

Recasens, A. R. C. (2016). *Where are they looking?* PhD thesis, Massachusetts Institute of Technology.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788.

Reichenberger, J., Pfaller, M., and Mühlberger, A. (2020). Gaze behavior in social fear conditioning: An eye-tracking study in virtual reality. *Frontiers in psychology*, 11:35.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99.

Reneker, J. C., Pannell, W. C., Babl, R. M., Zhang, Y., Lirette, S. T., Adah, F., and Reneker, M. R. (2020). Virtual immersive sensorimotor training (vist) in collegiate soccer athletes: A quasi-experimental study. *Heliyon*, 6(7):e04527.

Robertson, I. H., Manly, T., Andrade, J., Baddeley, B. T., and Yiend, J. (1997). Oops!': performance correlates of everyday attentional failures in traumatic brain injured and normal subjects. *Neuropsychologia*, 35(6):747–758.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520.

Schneider, M., Heine, A., Thaler, V., Torbeyns, J., De Smedt, B., Verschaffel, L., Jacobs, A. M., and Stern, E. (2008). A validation of eye movements as a measure of elementary school children's developing number sense. *Cognitive Development*, 23(3):409–422.

Smallwood, J., McSpadden, M., and Schooler, J. W. (2008). When attention matters: The curious incident of the wandering mind. *Memory & Cognition*, 36(6):1144–1150.

Sugano, Y., Matsushita, Y., and Sato, Y. (2012). Appearance-based gaze estimation using visual saliency. *IEEE transactions on pattern analysis and machine intelligence*, 35(2):329–341.

Tan, K.-H., Kriegman, D. J., and Ahuja, N. (2002). Appearance-based eye gaze estimation. In *Sixth IEEE Workshop on Applications of Computer Vision, 2002.(WACV 2002). Proceedings.*, pages 191–195. IEEE.

Vijayan, K. K., Mork, O. J., and Hansen, I. E. (2018). Eye tracker as a tool for engineering education. *Universal Journal of Educational Research*, 6(11):2647–2655.

Viola, P. and Jones, M. J. (2004). Robust real-time face detection. *International journal of computer vision*, 57(2):137–154.

Wanluk, N., Visitsattapongse, S., Juhong, A., and Pintavirooj, C. (2016). Smart wheelchair based on eye tracking. In *2016 9th Biomedical Engineering International Conference (BMEiCON)*, pages 1–4. IEEE.

Willemse, C. and Wykowska, A. (2019). In natural interaction with embodied robots, we prefer it when they follow our gaze: a gaze-contingent mobile eyetracking study. *Philosophical Transactions of the Royal Society B*, 374(1771):20180036.

Williams, O., Blake, A., and Cipolla, R. (2006). Sparse and semi-supervised visual mapping with the sˆ 3gp. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 230–237. IEEE.

Woolverton, G. A. and Pollastri, A. R. (2021). An exploration and critical examination of how "intelligent classroom technologies" can improve specific uses of direct student behavior observation methods. *Educational Measurement: Issues and Practice*.

Xie, R., Chen, Y., Wo, Y., and Wang, Q. (2019). A deep, information-theoretic framework for robust biometric recognition. *arXiv preprint arXiv:1902.08785*.

Yan, J., Lei, Z., Wen, L., and Li, S. Z. (2014). The fastest deformable part model for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2497–2504.

Yang, X. and Krajbich, I. (2021). Webcam-based online eye-tracking for behavioral research. *Judgment and Decision Making*, 16(6):1486.

Zandi, A. S., Quddus, A., Prest, L., and Comeau, F. J. (2019). Non-intrusive detection of drowsy driving based on eye tracking data. *Transportation research record*, 2673(6):247–257.

Zhang, C. and Zhang, Z. (2014). Improving multiview face detection with multi-task deep convolutional neural networks. In *IEEE Winter Conference on Applications of Computer Vision*, pages 1036–1041. IEEE.

Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503.

Zhang, X., Sugano, Y., Fritz, M., and Bulling, A. (2015). Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4511–4520.

Zhang, Z., Shen, W., Qiao, S., Wang, Y., Wang, B., and Yuille, A. (2020). Robust face detection via learning small faces on hard images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1361–1370.